DOCUMENT RESUME

ED 129 866 TM 005 663

AUTHOR Livingston, Samuel A.
TITLE The Criterion-Referenced Reliability of a Single Score. Report 76-01.
INSTITUTION Educational Testing Service, Princeton, N.J. Center for Occupational and Professional Assessment.
REPORT NO ETS-COPA-76-01
PUB DATE Apr 76
NOTE 11p.

EDRS PRICE MF-$0.83 HC-$1.67 Plus Postage.
DESCRIPTORS *Classification; *Criterion Referenced Tests; *Measurement; *Probability; Reliability; Scores; Standard Error of Measurement; Test Interpretation; *Test Reliability; True Scores

ABSTRACT

A distinction is made between reliability of measurement and reliability of classification; the "criterion-referenced reliability coefficient" describes the former. Application of this coefficient to the probability distribution of possible scores for a single student yields a meaningful way to describe the reliability of a single score. (Author)

****
* Documents acquired by ERIC include many informal unpublished *
* materials not available from other sources. ERIC makes every effort *
* to obtain the best copy available. Nevertheless, items of marginal *
* reproducibility are often encountered and this affects the quality *
* of the microfiche and hardcopy reproductions ERIC makes available *
* via the ERIC Document Reproduction Service (EDRS). EDRS is not *
* responsible for the quality of the original document. Reproductions *
* supplied by EDRS are the best that can be made from the original. *
****

ED129866

COPA-76-01

The Criterion-Referenced Reliability of a Single Score

Samuel A. Livingston

Center for Occupational and Professional
Assessment
Educational Testing Service

Princeton, New Jersey

April, 1976

This Bulletin is a draft for interoffice circulation. Corrections and suggestions for revision are solicited. The Bulletin should not be cited as a reference without the specific permission of the author. It is automatically superseded upon formal publication of the material.

TM 005 663

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

"PERMISSION TO REPRODUCE THIS COPYRIGHTED MATERIAL HAS BEEN GRANTED BY

SAMUEL A. LIVINGSTON

TO ERIC AND ORGANIZATIONS OPERATING UNDER AGREEMENTS WITH THE NATIONAL INSTITUTE OF EDUCATION. FURTHER REPRODUCTION OUTSIDE THE ERIC SYSTEM REQUIRES PERMISSION OF THE COPYRIGHT OWNER."

The Criterion-Referenced Reliability of a Single Score

ABSTRACT

A distinction is made between reliability of measurement and reliability of classification; the "criterion-referenced reliability coefficient" describes the former. Application of this coefficient to the probability distribution of possible scores for a single student yields a meaningful way to describe the reliability of a single score.

The Criterion-Referenced Reliability of a Single Score

In recent years there has been some theoretical controversy concerning the concept of reliability for criterion-referenced tests. Much of the controversy has been of the blind-men-and-elephant type, in that the participants seem to have been unaware that they were not all talking about the same thing. The key distinction that seems to have been overlooked is the distinction between the reliability of measurement and the reliability of classification. Measurement is the process of assigning numbers to things in some meaningful way. Classification is the process of sorting things into groups. Of course, a very common method of classification is to specify a rule based on a measurement, as when (for example) students with test scores of 95 or higher are classified as "masters" while those with test scores below 95 are classified as "non-masters". Nevertheless, the processes of measurement and classification are at least conceptually distinct. It is meaningful to talk about the reliability of the classification process, and it is also meaningful to talk about the reliability of the measurement process without reference to any classification that may be made on the basis of the measurement. Hambleton and Novick (1973), Swaminathan, et al (1974), Marshall and Haertel (1975), and Fricke (1972) have all dealt with the problems of the reliability of classification. The reliability of measurement is the subject of classical test theory, and this theory has been extended by Livingston (1972) to apply to situations where measurement is made with reference to points other than the mean of a score distribution. (Such a point might represent, for example,

a "baseline", or a "target level" for training, or the level of success expected by pure chance in the absence of any real knowledge.) The purpose of the present paper is to extend the concept of the reliability of measurement still further, to include not only the reliability of a group of scores, but also the reliability of a single score.

When we make a measurement in education, we are interested in generalizing beyond the single observation we have made. Many conditions that are irrelevant to instructional purposes may influence a student's test score: the particular questions included in the test (from among all the questions that might have been included), the format and phrasing of the questions, how much sleep the student had the previous night, whether he had an argument with his friend that morning, and so on. Therefore we recognize that the score the student obtained is only one of several different scores he could possibly have obtained, even with the instructionally relevant conditions completely specified. Associated with each of these possible scores is a probability, and these probabilities form a probability distribution. (Lord and Novick, 1968, p. 30, refer to this distribution as a "propensity distribution".) The mean of this probability distribution is the student's "true score", and the standard deviation of this probability distribution is the standard error of measurement. Thus, although in practice the standard error of measurement is usually estimated from the scores of several students, conceptually it is a characteristic of the score of a single student.

Now suppose it were possible for a student to take a test twice, with no change in the instructionally relevant conditions, but with the possibility of random variation in the irrelevant conditions. Associated with each possible pair of scores the student might obtain is a probability, and these

probabilities form a joint probability distribution. Since the only sources of variation in either variable are the instructionally irrelevant conditions that have been assumed to vary at random, this joint probability distribution is characterized by a correlation coefficient of zero. However, if our reference point for measurement is some point other than the student's true score, we might be interested in a product-moment correlation based on deviations from the reference point, rather than from the mean, i.e., the "criterion-referenced correlation coefficient" (Livingston, 1972, pp. 14-15). This coefficient is defined by the formula

$$\frac{\mathcal{E}(X - C_x)(Y - C_y)}{\sqrt{\mathcal{E}(X - C_x)^2 \; \mathcal{E}(Y - C_y)^2}}$$

where X and Y are the two variables, C_x and C_y are their reference points, and $\mathcal{E}$ indicates the expected value. It can be shown (Livingston, 1972, p. 17) that the above expression is equivalent to

$$\frac{r\, s_x\, s_y + (m_x - C_x)(m_y - C_y)}{\sqrt{[s_x^2 + (m_x - C_x)^2]\,[s_y^2 + (m_y - C_y)^2]}}$$

where m_x and m_y are the means of the variables, s_x and s_y are their standard deviations, and r is the conventional product-moment correlation between them. Since the two variables have identical distributions and the same reference point, the formula for the criterion-referenced correlation coefficient reduces to

$$\frac{r\, s^2 + (m - C)^2}{s^2 + (m - C)^2} \quad .$$

But in this joint probability distribution of possible scores for a single student taking the test twice, r is zero, m is the student's true score, and s is the standard error of measurement. The formula thus becomes

$$\frac{(T - C)^2}{s^2(E) + (T - C)^2} \quad .$$

where T is the student's true score and $s^2(E)$ is his standard error of measurement for the test.

And, since the two variables in the joint probability distribution are scores on the same test for the same student, with all differences being the result of random variation in instructionally irrelevant conditions -- i.e., measurement error -- we have a meaningful way to define a reliability coefficient for a single observation.

The same formula can be arrived at in another way, by expressing the criterion-referenced reliability coefficient as

$$\frac{s^2(T) + (m - C)^2}{s^2(T) + (m - C)^2 + s^2(E)} \quad ,$$

where $s^2(T)$ is the variance of true scores (Livingston, 1972, p. 19). Since we are dealing with the probability distribution of possible scores for a single student, the variance of true scores is zero, and m, the mean of the distribution, is the student's true score. Making these substitutions, we again have

$$\frac{(T - C)^2}{(T - C)^2 + s^2(E)} \quad .$$

Notice that the reliability of a single score, as defined here, is a simple function of the ratio of the difference between the student's true score and the reference point (T - C) to the standard error of measurement s(E). If h represents the ratio (T - C)/s(E), then the reliability of the score is $h^2/(h^2+1)$. For example, if the student's true score is one standard error away from the reference point, the reliability of the observed score is 1/2, or .50. If the difference is three standard errors, the reliability is 9/10, or .90. If the student's true score exactly equals the reference point, the reliability of the observed score is zero.

But what do these numbers mean? What does it mean to say that the reliability of a student's observed score on a test is .90? It means that 90 per cent of the information conveyed by the student's observed score is reliable information. This interpretation is analogous to a similar interpretation we might make for a conventional norm-referenced reliability coefficient in a group of students. Since r is the ratio of "true" variance to observed variance, we might say that r is the proportion of the information given by the group of scores that is reliable information. Since the purpose of a norm-referenced test is to discriminate between students, the variance of scores is a meaningful measure of the information they provide. For a single student whose score is to be compared with a specified reference point, the analogous measure is the squared difference between his score and the reference point.

How, then, can we estimate this quantity? There are at least two possible methods. If we approach the problem from the item-sampling viewpoint (as do Kriewall, 1969, and Millman, 1972), we can estimate T by the number of items answered correctly and estimate $s^2(E)$ by $np(1 - p)$, where n is the number of items and p is the proportion of items answered correctly.

Note that the criterion-referenced reliability of a single score is not an estimate of the probability of correctly classifying a student as above or below the reference point. This probability, for a student whose true score is three standard errors from the reference point, would not be .90, but .9987, assuming that errors of measurement are normally distributed.

Alternatively, we can use data from the whole group of examinees to estimate T and $s^2(E)$. In this method we would use the estimates

$$T = m + r(X - m)$$

$$s^2(E) = (1 - r)s^2$$

where X is the student's observed score, m is the group mean, s^2 is the variance of observed scores in the group, and r is the conventional reliability coefficient computed from the scores of the whole group.

Acknowledgements

I wish to thank Ronald K. Hambleton, John R. Hills, Ollie A. Jensen, Frederic M. Lord, Jason Millman, Michael M. Ravitch, and Cheryl W. Reed for their many helpful comments on earlier drafts of this paper.

REFERENCES

FRICKE, R. Testgütekriterien bei lehrzielorientierten Tests. Zeitschrift fur Erziehungswissenschaftliche Forschung, 1972, 6, 150-175.

HAMBLETON, R. K. and NOVICK, M. R. Toward an integration of theory and method for criterion-referenced tests. Journal of Educational Measurement, 1973, 10, 159-170.

KRIEWALL, T. E. Applications of information theory and acceptance sampling principles to the management of mathematics instruction. Unpublished doctoral dissertation, University of Wisconsin, 1969.

LIVINGSTON, S. A. Criterion-referenced applications of classical test theory. Journal of Educational Measurement, 1972, 9, 13-26.

LORD, F. M. and NOVICK, M. R. Statistical Theories of Mental Test Scores. Reading, Massachusetts: Addison-Wesley, 1968.

MARSHALL, J. L. and HAERTEL, E. H. A single-administration reliability index for criterion-referenced tests: the mean split-half coefficient of agreement. Madison, Wisconsin: University of Wisconsin Research and Development Center for Cognitive Learning, 1975.

MILLMAN, J. Tables needed for determining number of items needed on domain-referenced tests and number of students to be tested. Technical Paper No. 6. Los Angeles: Instructional Objectives Exchange, 1972.

SWAMINATHAN, H., HAMBLETON, R. K., and ALGINA, J. J. Reliability of criterion-referenced tests: a decision-theoretic formulation. Journal of Educational Measurement, 1974, 11, 263-267.